Mathématiques II
HEC Lausanne, Spring 2025

Fiche de théorie 11

Linéarité du gradient et du hessien et règle de chaîne

Soit $f, g : \mathbb{R}^n \to \mathbb{R}$ des fonctions et soit $H_f(\boldsymbol{x})$ désigner son Hessien.

Puisque la dérivée est linéaire ($(f+g)' = f' + g'$), nous avons la règle suivante, qui simplifie la recherche de gradients et de Hessians :

$$\boxed{\begin{array}{c}\operatorname{grad}(f(\boldsymbol{x}) + g(\boldsymbol{x})) = \operatorname{grad} f(\boldsymbol{x}) + \operatorname{grad} g(\boldsymbol{x}) \\ H_{f+g}(\boldsymbol{x}) = H_f(\boldsymbol{x}) + H_g(\boldsymbol{x})\end{array}}$$

Si $h : \mathbb{R} \to \mathbb{R}$ est une autre fonction et que nous considérons la composition $h(f(\boldsymbol{x}))$, alors par la règle de chaîne habituelle $(h(f(x)))' = h'(f(x))\, f'(x)$ nous avons que la même chose vaut pour le gradient :

$$\boxed{\operatorname{grad} h(f(\boldsymbol{x})) = h'(f(\boldsymbol{x}))\ \operatorname{grad} f(\boldsymbol{x}).}$$

Ces règles sont simples, mais très utiles.

Exemple 1. $\operatorname{grad}(f(\boldsymbol{x}))^4 = 4f(\boldsymbol{x})^3\ \operatorname{grad} f(\boldsymbol{x})$.

Gradient et Hessien d'une fonction quadratique

Considérons $A \in M_{2,2}$ symétrique, $\boldsymbol{b} \in \mathbb{R}^2$ et la quadratique fonction

$$f(x_1, x_2) = \boldsymbol{x}^\top A\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} + c = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 + b_1x_1 + b_2x_2 + c.$$

Alors

$$\operatorname{grad} f(x_1, x_2) = \begin{pmatrix} f'_{x_1} \\ f'_{x_2} \end{pmatrix} = \begin{pmatrix} 2a_{11}x_1 + 2a_{12}x_2 + b_1 \\ 2a_{12}x_1 + 2a_{22}x_2 + b_2 \end{pmatrix}.$$

Notons que ceci peut être écrit comme

$$\operatorname{grad} f(x_1, x_2) = 2\begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = 2A\boldsymbol{x} + \boldsymbol{b}.$$

En fait, les mêmes formules valent dans les dimensions supérieures :

Théorème 1. *Si* $A \in M_{n,n}$ *et* $\boldsymbol{b} \in \mathbb{R}^n$ *, nous avons*

$$\operatorname{grad} \boldsymbol{x}^\top A\boldsymbol{x} = 2A\boldsymbol{x} \quad \textit{and} \quad \operatorname{grad} \boldsymbol{b}^\top \boldsymbol{x} = \boldsymbol{b}.$$

Proof. We have

$$\operatorname{grad} \boldsymbol{x}^\top A\boldsymbol{x} = \operatorname{grad} \sum_{i,j=1}^{n} a_{ij} x_i x_j = \sum_{i,j=1}^{n} a_{ij} \operatorname{grad} x_i x_j,$$

so we only need to calculate $\operatorname{grad} x_i x_j$:

$$\operatorname{grad} x_i x_j = x_i\, \boldsymbol{e}_j + x_j\, \boldsymbol{e}_i,$$

where $\boldsymbol{e}_i$ is a vector of zeroes with 1 at i^{th} row. Hence,

$$\operatorname{grad} \boldsymbol{x}^\top A\boldsymbol{x} = \sum_{i,j=1}^{n} a_{ij}(x_i\, \boldsymbol{e}_j + x_j\, \boldsymbol{e}_i) = \sum_{i,j=1}^{n} a_{ij}\, x_i\, \boldsymbol{e}_j + \sum_{i,j=1}^{n} a_{ij}\, x_j\, \boldsymbol{e}_i.$$

Since $a_{ij} = a_{ji}$ by symmetry of A, we have

$$\sum_{i,j=1}^{n} a_{ji}\, x_i\, \boldsymbol{e}_j = \sum_{i,j=1}^{n} a_{ij}\, x_j\, \boldsymbol{e}_i = A\boldsymbol{x},$$

which concludes the proof of the first claim. The second claim follows similarly:

$$\operatorname{grad} \boldsymbol{b}^\top \boldsymbol{x} = \operatorname{grad} \sum_{i=1}^{n} b_i x_i = \sum_{i=1}^{n} b_i \operatorname{grad} x_i = \sum_{i=1}^{n} b_i \boldsymbol{e}_i = \boldsymbol{b},$$

where we used that $\operatorname{grad} x_i = \boldsymbol{e}_i$. □

Donc,

$$\boxed{\operatorname{grad}(\boldsymbol{x}^\top A\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} + c) = 2A\boldsymbol{x} + \boldsymbol{b}.}$$

Nous allons utiliser ceci plus tard.

Remarque 1. *Compare this with* $\frac{d}{dx}(ax^2 + bx + c) = 2ax + b$.

Ensuite, nous voulons calculer le Hessien de la fonction quadratique. Puisque le Hessien est linéaire et les dérivées secondes de $\boldsymbol{b}^\top \boldsymbol{x} + c$ sont clairement nulles, nous avons seulement besoin de trouver le Hessien de $\boldsymbol{x}^\top A\boldsymbol{x}$. UNE calcul similaire donne

$$\boxed{H_f(\boldsymbol{x}) = 2A.}$$

Par conséquent, nous savons comment trouver à la fois le gradient et le Hessien d'une fonction !

Régression linéaire

Considérons le problème suivant : nous avons des points x_i et y_i , $i = 1, \ldots, n$ obtenus à partir de certaines données. Nous traçons ces points et voyons qu'ils se situent près d'une droite. Comment trouver la meilleure droite qui correspond à ces points ?

Une droite sur le plan peut être exprimée par l'équation suivante :

$$y = ax + b.$$

Nous ne pouvons pas simplement supposer que $y_i = ax_i + b$ pour tout $i = 1, \dots, n$ et essayer de résoudre a, b, car ce serait trop d'équations. Autrement dit, à moins qu'ils ne se situent tous exactement sur la même ligne pour commencer.

Ce que nous pouvons faire à la place est de supposer que $y_i = ax_i + b + \varepsilon_i$, où ε_i sont des erreurs. Maintenant, nous pouvons trouver a et b, qui minimisent l'erreur totale.

Mais qu'entendons-nous par erreur totale ? Il existe de nombreuses façons d'y répondre, le plus standard est donné par la somme des erreurs au carré :

$$R = \sum_{i=1}^{n} \varepsilon_i^2.$$

En branchant $\varepsilon_i = y_i - ax_i - b$ dans cette formule, nous voyons que R est une fonction de a et b :

$$\boxed{R(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2.}$$

Nous pouvons maintenant minimiser cette fonction par rapport à a et b et obtenir la droite dite de régression des moindres carrés.

Remarque 2. *Ceci est de loin la seule façon de définir ce que " meilleure ligne " signifie dans ce contexte. Une autre alternative tout aussi intéressante consiste à minimiser*

$$L(a, b) = \sum_{i=1}^{n} |\varepsilon_i| = \sum_{i=1}^{n} |y_i - ax_i - b|.$$

Ceci est connu comme régression linéaire robuste. Malheureusement, $|x|$ est non différentiable, donc la minimisation dans ce cas devient plus complexe.

Pour minimiser R, nous calculons d'abord son gradient :

$$\operatorname{grad} R(a, b) = \begin{pmatrix} R_a \\ R_b \end{pmatrix} = \begin{pmatrix} -2\sum_{i=1}^{n} (y_i - ax_i - b)x_i \\ -2\sum_{i=1}^{n} (y_i - ax_i - b) \end{pmatrix},$$

égalons-le à zéro et obtenons deux équations :

$$\sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - bn = 0.$$

En notant

$$\rho = \frac{1}{n}\sum_{i=1}^{n} y_i x_i, \quad s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad \mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

nous trouvons que les équations ci-dessus peuvent être réécrites comme

$$\begin{cases} \rho - as^2 - b\mu_x = 0, \\ \mu_y - a\mu_x - b = 0. \end{cases}$$

En résolvant ce système, nous obtenons

$$\boxed{a = \frac{\rho - \mu_x \mu_y}{s^2 - \mu_x^2}, \quad b = \mu_y + a\mu_x.}$$

Si nous branchons $\rho,\ \mu_x,\ \mu_y$ et s^2, nous obtiendrons

$$a = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i x_i - \frac{1}{n^2}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}$$

Nous pouvons aussi remanier un peu cette formule et la réécrire sous une autre forme :

$$a = \frac{\displaystyle\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{\displaystyle\sum_{i=1}^{n} (x_i - \mu_x)^2}.$$

Rappelons toutefois qu'il s'agit de points candidats, et nous devons vérifier qu'ils minimisent effectivement R. À cette fin, nous calculons le Hessien de R :

$$H_R(\boldsymbol{x}) = \begin{pmatrix} 2\sum_{i=1}^{n} x_i^2 & 2\sum_{i=1} x_i \\ 2\sum_{i=1}^{n} x_i & 2n \end{pmatrix}.$$

Clairement,

$$2\sum_{i=1}^{n} x_i^2 > 0,$$

nous devons donc vérifier le déterminant et appliquer la loi d'inertie de Sylvester. Nous avons

$$\begin{aligned} \det H_R(\boldsymbol{x}) &= 4n\sum_{i=1}^{n} x_i^2 - 4\left(\sum_{i=1}^{n} x_i\right)^2 \\ &= 4n^2\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n^2}\sum_{i=1}^{n} x_i\right)^2\right) \\ &= 4n\sum_{i=1}^{n} (x_i - \mu_x)^2 > 0. \end{aligned}$$

Remarque 3. *Ci-dessus, nous avons utilisé le calcul suivant :*

$$\begin{aligned}\sum_i (x_i - \mu_x)^2 &= \sum_i (x_i^2 - 2x_i\mu_x + \mu_x^2) \\ &= \sum_i x_i^2 - 2\mu_x \sum_i x_i + n\mu_x^2 \\ &= \sum_i x_i^2 - 2n\mu_x^2 + n\mu_x^2 \\ &= \sum_i x_i^2 - n\mu_x^2.\end{aligned}$$

Régression linéaire sous forme matricielle

Résolvons le même problème en utilisant la notation matricielle. Nous verrons alors que cette approche permet de réaliser des types de régression beaucoup plus généraux !

Soit y_i et x_i, $i = 1, \ldots, n$ nos points de données. Notons

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} b \\ a \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \boldsymbol{y} - X\boldsymbol{\beta}.$$

Notons que X est une matrice $n \times 2$. La même fonction R comme ci-dessus dans cette notation peut être écrite comme

$$R(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}.$$

En branchant la définition de $\boldsymbol{\varepsilon}$, nous obtenons

$$R(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{\beta}^\top X^\top \boldsymbol{y} - \boldsymbol{y}^\top X\boldsymbol{\beta} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}.$$

Puisque $\boldsymbol{y}^\top X\boldsymbol{\beta} = \boldsymbol{\beta}^\top X^\top \boldsymbol{y}$ (nous pouvons toujours transposer un nombre !), nous avons

$$R(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top A\boldsymbol{\beta} - 2\,\boldsymbol{b}^\top \boldsymbol{\beta} + c,$$

où $A = X^\top X$, $\boldsymbol{b} = X^\top \boldsymbol{y}$ et $c = \boldsymbol{y}^\top \boldsymbol{y}$.

Selon le théorème au début de ce cours, le gradient de R est donné par

$$\operatorname{grad} R(\boldsymbol{\beta}) = 2A\boldsymbol{\beta} - 2\boldsymbol{b}.$$

En l'égalant à zéro, nous constatons que $\boldsymbol{\beta}$ satisfait

$$A\boldsymbol{\beta} - \boldsymbol{b} = \boldsymbol{0}.$$

Donc,

$$\boldsymbol{\beta} = A^{-1}\boldsymbol{b}.$$

Il reste à brancher les formules pour A et $\boldsymbol{b}$ pour obtenir

$$\boxed{\boldsymbol{\beta} = (X^\top X)^{-1} X^\top \boldsymbol{y}.}$$

Cette formule est appelée l'équation normale de la régression linéaire, elle donne les coefficients de régression linéaire directement en fonction des données d'origine points !

Remarque 4. *Note that if* $\boldsymbol{y} \in \mathbb{R}^n$ *and* $X \in M_{n,2}$, *then* $X^\top \in M_{2,n}$, *hence* $X^\top \boldsymbol{y} \in \mathbb{R}^2$ *and* $X^\top X \in M_{2,2}$, *so all products are well-defined.*

The matrix $X^\top X$ *is strictly positive definite, which means that it does not have zero eigenvalues, which means that it is invertible!*

Remarque 5 (Important)**.** *Note that we cannot simplify* $(X^\top X)^{-1} X^\top$ *using* $(AB)^{-1} = B^{-1}A^{-1}$. *Why? Because the last formula is only true if* A *and* B *are invertible. In particular, they must be square! Note that although* $X^\top X$ *is a square matrix* (2×2), X *and* $X^\top$ *are not (they are* $n \times 2$ *and* $2 \times n$ *correspondingly). Hence, their inverses do not even make sense!*

Comment utiliser l'équation normale en pratique ?

Après avoir dérivé l'équation normale, nous pouvons oublier sa dérivation et simplement l'utiliser comme suit :

- Étant donné les points de données x_i, y_i, $i = 1, \ldots, n$, construire un vecteur $\boldsymbol{y}$ et une matrice X comme ci-dessus.
- Calculer $\boldsymbol{\beta} = (X^\top X)^{-1} X^\top \boldsymbol{y}$.
- Maintenant, la meilleure droite qui approche nos points de données est donnée par $y = \beta_1 + \beta_2 x$.

Régression linéaire généralisée

Que faire si nous voulons trouver la meilleure courbe, au lieu d'une droite, qui approxime notre ensemble de données x_i, y_i ? Par exemple, nous pouvons trouver une courbe polynomiale approximant nos données :

$$y_i \approx \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p$$

en utilisant la même idée ! À cette fin, nous construisons l'erreur d'approximation

$$R(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \cdots - \beta_p x_i^p)^2$$

et minimisons-le par rapport à $\boldsymbol{\beta}$. Ceci devient beaucoup plus facile dans la notation matricielle. Soit

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^p \\ 1 & x_2 & x_2^2 & \ldots & x_2^p \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_n & x_n^2 & \ldots & x_n^p \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

Avec cette notation, R devient

$$R(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}).$$

Notons que le problème est nettement plus général (polynomial au lieu de droite), mais la fonction que nous devons optimiser est exactement la même !

Par conséquent, nous pouvons immédiatement écrire la solution de ce problème de minimisation, c'est-à-dire, l'équation normale :

$$\boldsymbol{\beta} = (X^\top X)^{-1} X^\top \boldsymbol{y}.$$

Cette équation nous indique maintenant que la meilleure courbe polynomiale qui approxime nos données est donnée par

$$y = \beta_0 + \beta_1 x + \cdots + \beta_p x^p.$$

Remarque 6. *In fact, there is nothing special about polynomials. We can approximate our data by a linear combination of any functions:*

$$y_i \approx \sum_{k=1}^{p} \beta_k \varphi_k(x_i).$$

To this end, we just change the definition of the matrix X to

$$X = \begin{pmatrix} \varphi_0(x_1) & \varphi_1(x_1) & \dots & \varphi_p(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) & \dots & \varphi_p(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \dots & \varphi_p(x_n) \end{pmatrix},$$

and use the same normal equation with this new matrix X. One neat example is to approximate seasonal data with periodic functions:

$$y_i \approx \beta_0 + \beta_1 \cos\left(\frac{x_i}{T_1}\right) + \beta_2 \cos\left(\frac{x_i}{T_2}\right),$$

where x_i is interpreted as time and T_1, T_2 are two different time scales.

Une autre extension naturelle

Une autre extension naturelle de la méthode précédente se produit lorsque nous voulons modéliser l'effet de nombreux facteurs x_i, z_i, t_i, ... sur y_i. Géométriquement, cela signifie que nous essayons de trouver le meilleur hyperplan pour approximer nos points.

Supposons que $y_i \approx \beta_0 + \beta_1 x_i + \beta_2 z_i$ (seulement deux facteurs). Alors l'erreur de cette approximation est

$$R(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2,$$

qui peut à nouveau être écrite de la même manière avec la seule différence que X est maintenant

$$X = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & z_n \end{pmatrix}.$$

Tout le reste dans le modèle reste le même, donc le $\boldsymbol{\beta}$ résultant est à nouveau donné par l'équation normale.